Structural knowledge for decomposing image sequences

Gregory J. Power
Air Force Research Laboratory
Target Recognition Branch, AFRL/SNAT,
Building 620, 2241 Avionics Circle
Wright-Patterson AFB, Ohio 45433-7321

Abstract ***The storage and retrieval of digital image sequences is becoming more important with the proliferation of computers, the internet, digital TV and digital cameras. This paper suggests using a prioi knowledge in the design to improve the decomposition of digital image sequences for data storage and retrieval. In particular, the paper focuses on the specific image sequence structure which differs depending on the application. Different image sequence structures dictate different image sequence decomposition approaches. Examples of various image sequence structures are given. The structural knowledge of the image sequence is shown to impact the video analysis tools used for decomposition.***

Keywords: Image Sequence Analysis, Data Storage and Retrieval, Motion Imagery, Digital Video, video on demand

1 Introduction

Digital image sequences imply a discrete sequence of two-dimensional discrete images. Digital image sequences are also known by other names. Some textbooks use the term digital video [1, 2]. The National Image Mapping Agency (NIMA) includes digital image sequences under the project heading of "motion imagery". In this paper, I will use the term digital video and digital image sequence interchangeably. The digital image sequences are produced for a variety of purposes including entertainment, educational, medical and military. There is an overwhelming amount of digital video with the advent and proliferation of digital information through computers and the internet along with the trend toward digitizing applications that were formerly analog including TV, cameras, and cinema. Along with the increase in digital video, is an increase demand for storing and retrieving the digital video. So, in recent years many techniques have been researched to address the problem of storing and retrieving. Video analysis plays an important role in the video storage and retrieval process. It is used to find cuts, fades, pans, action shots, key frames, and activity, for example. This paper suggests a key consideration that can improve the storing, indexing and retrieving of video. The improvement is achieved using *a priori* knowledge about the application

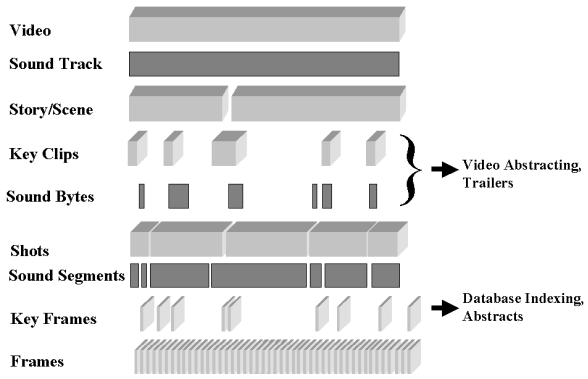

Figure 1: A cinematic film structure model consists of stories or scenes, sound track, key clips, sound bytes, shots, and key frames.

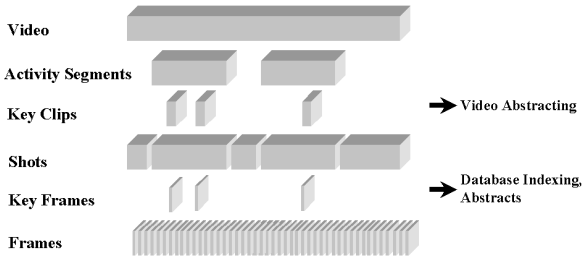

Figure 2: Surveillance video structure can be thought of as consisting of activity segments, key clips, shots, and key frames.

whether it is cinematic, military sensor, security surveillance, or another video type. Each application is shown to have its own inherent structure that can assist the video analysis and decomposition.

2 Structural Analysis

For the public consumer, video across the world wide web (WWW) and video-on-demand (VOD) is a driving force behind the use of digital image sequence analysis. The public VOD requirements include cinematic and TV for entertainment purposes as well as VOD for consumer purchasing, instructional, medical, and other personal research applications. For TV and cinematic film, the video can be decomposed as shown in Figure 1. Knowledge of the TV and cinematic rules for editing as well as incorporating additional information from the sound track can assist in abstracting the video.

VOD systems that are built solely for the public consumer will not meet all the requirements of specialized consumers such as medical, military, private security and governmental consumers. There are differences in the types of digital video that will be stored and retrieved. Customized systems can take advantage of the differing need. For instance, the shots from a movie trailer may be made up of scene clips with abrupt changes and fades that are constructed based on specialized movie trailer editing rules. However, a video surveillance camera which is fixed and pointing at a hallway

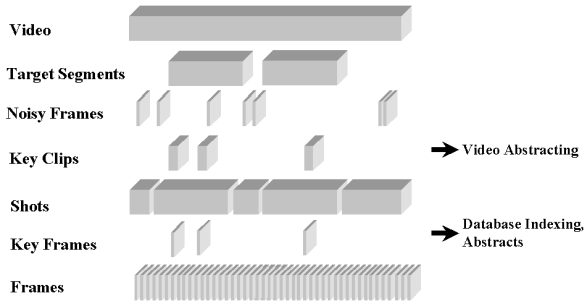

Figure 3: A raw air-to-ground infrared video structure can be thought of as consisting of target segments, erratic noise frames, key clips, shots and key frames. The shots may be triggered by an operator, navigation, weapon or other systems.

may only be made up of clips that start and end when some continuous motion is detected (Figure 2). Therefore, by comparing Figure 1 and Figure 2, it is obvious that different video structuring needs different analysis techniques to create the key frames and shots needed for storage and retrieval.

Again, the impact of video structure is evident by comparing Figures 1 and 2 with yet another application model shown in Figure 3 which shows an air-to-ground infrared video structure model. Instead of scenes, the infrared video has target segments. In addition, it may have noise frames. It may have shot boundaries that are not due to abrupt changes but those shot boundaries may be labeled based on a significant event rather than a large dynamic image change. For instance, a significant event could be triggered by instruments that indicate the aircraft has arrived over a particular geographical coordinate; or the aircraft has detected a radar, detected a ground vehicle movement, acquired a target, or fired a weapon.

3 Structures Impact on Video Analysis

Structural analysis suggests that video analysis techniques must be customized for the partic-

ular video task. For instance, a video analysis technique that detects abrupt changes is good for movie sequences and TV where high quality is guaranteed, and an abrupt change signals a shot change. However, for a sensor on board an aircraft that is slowly mapping the ground, there should be no abrupt change, so an abrupt change may signal erratic noise. For this aircraft sensor, collateral information from instruments instead of information from video analysis, is needed to determine the significant video shots. So, the use of a particular video analysis technique to find abrupt changes is used differently depending on the application.

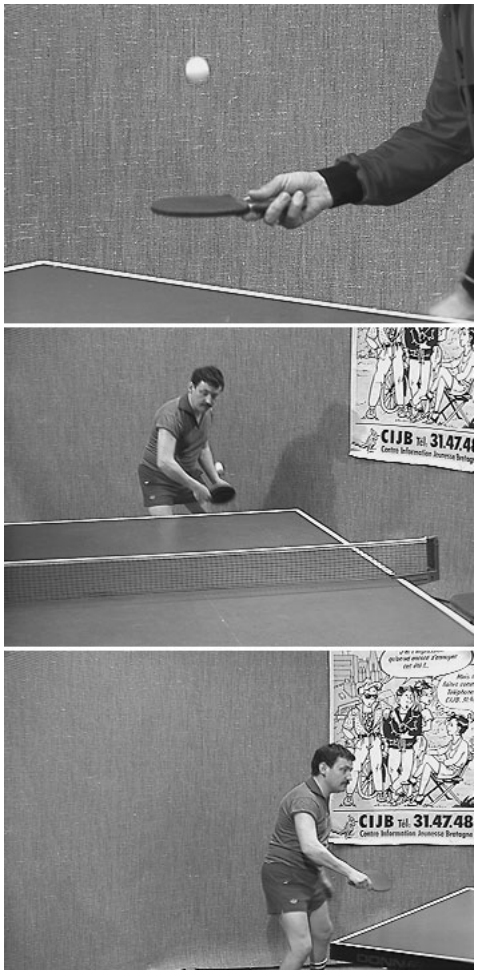

Figure 4: Key frames from the three scenes in the table tennis image sequence.

3.1 The velocital information feature analysis example

As an example, this section uses the velocital information feature [3] as a video analysis technique on a TV-like sequence and an air-to-ground infrared sequence. The velocital information feature has been found to be useful for detecting sudden changes. Assuming a total of P pixels in a frame and given the velocity at a particular pixel as $VI[x, y, t_n]$ then the velocital information feature, $VI_{stdev}[t_n]$, for an image frame can be estimated by calculating the standard deviation as

$$VI_{stdev}[t_n] =$$

$$\sqrt{\left[\frac{1}{P}\sum_x\sum_y VI^2(x,y,t_n)\right] - VI^2_{mean}(t_n)},$$

in which

$$VI_{mean}(t_n) = \frac{1}{P}\sum_x\sum_y VI(x,y,t_n).$$

The TV-like sequence chosen for this example is a portion of the standard table tennis sequence which has three shots. Frames from each of the shots are shown in Figure 4. Sample images from the air-to-ground image sequence is shown in Figure 5 which shows one clean frame and one frame with erratic noise.

The velocital information feature is plotted for both the table tennis and the air-to-ground infrared image sequence. The resulting plots are shown in Figure 6. The sudden changes are noted by the peaks. For the table tennis sequence, each new shot is correctly detected showing the three distinct shots illustrated in Figure 4. For the infrared sequence which was only one continuous video with no shots, the noisy frame number 14 was correctly selected by the velocital information feature. This velocital information feature example demonstrates that the results from the same video analysis tool detects different events (i.e. shots or noise) on video clips that have different underlying video structures.

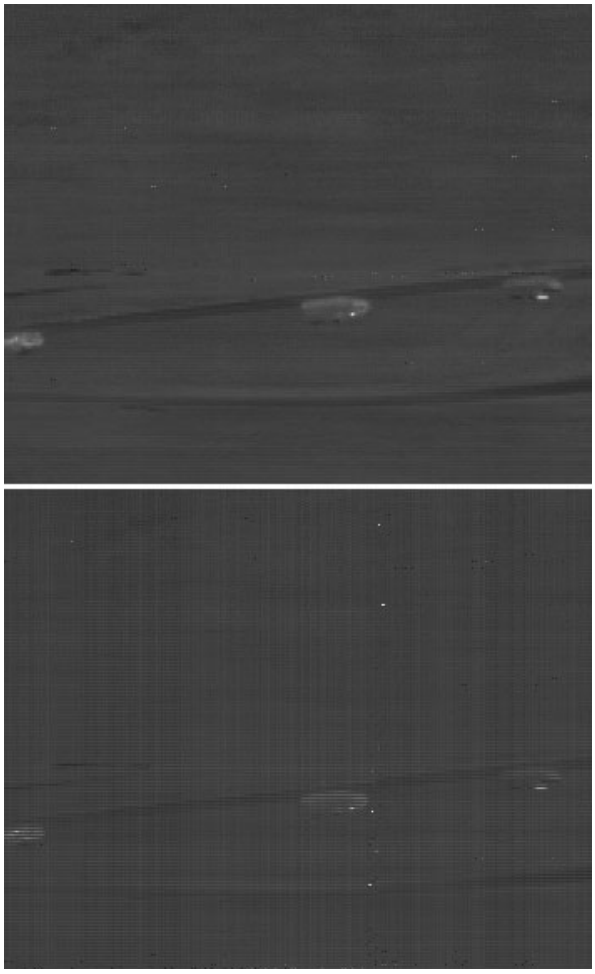

Figure 5: Key frames from an air-to-ground infrared image sequence. One clean frame (above) and one frame with erratic noise (below).

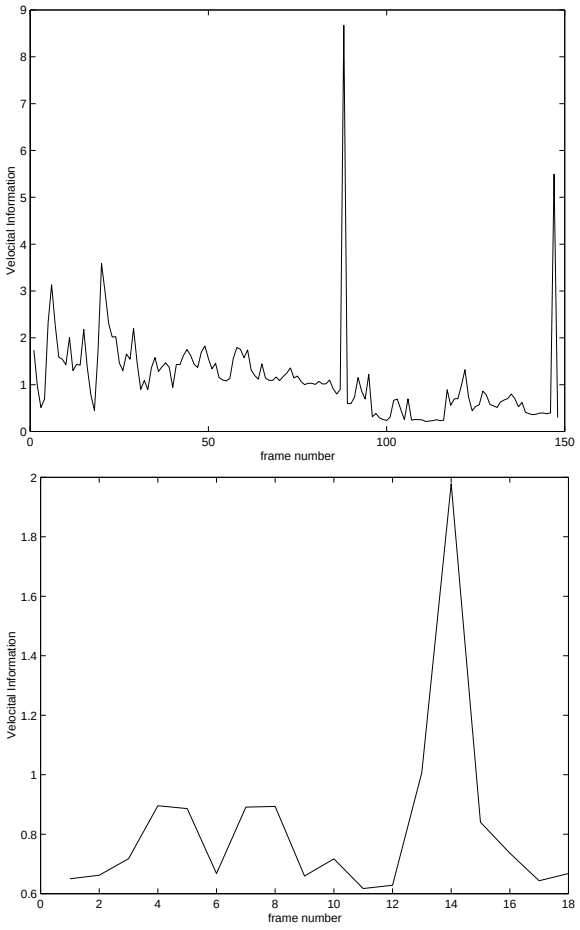

Figure 6: Velocital information for the 149-frame table tennis sequence (above) and an 18-frame infrared image sequence (below). Sudden changes for the table tennis sequence are shot boundaries but for the infrared sequence sudden changes represent erratic noise.

3.2 The impact of variations within classes

When considering the video structure, one more point to keep in mind is that the three video structures shown in Figures 1, 2, and 3 are not representative of all the videos within their class. For instance, consider the fixed surveillance camera structure model. If a hallway has a fixed camera with a fixed light source and no strange dynamic obscurations (such as an open door allowing the wind to blow leaves around all day long), then the model applies. However, consider a variation of the fixed surveillance camera such as the fixed camera pointing at the orangutan play pen at the Washington zoo. A few sample images are shown in Figure 7. Even without activity within the play pen, a variety of dynamic factors change the resulting image that is transmitted across the web including reflections on the glass from passers-by, changes in sunlight and shadow, low sampling rate and compression. Using the velocital information feature to chart the dynamic changes on a sequence of play pen frames results in the plot shown in Figure 8. The velocital information

Figure 7: Web camera fixed on the orangutan play pen at the Washington Zoo. Other than the presence of the orangutan, variations in image scenes are due to lighting, glass reflections, and compression.

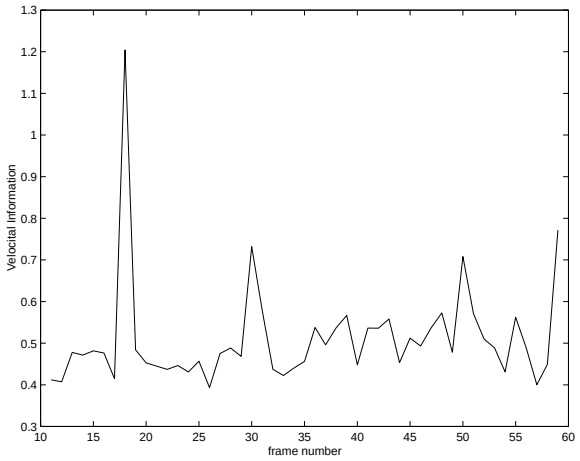

Figure 8: Plot of dynamic changes in a sequence of frames from the the orangutan play pen. Significant events occur at frames 18, 30, 50, and 59.

does pick up significant changes (Figure 9) but it also has a higher value for null periods due to the variations within the class. These variations can cause the shot detector to miss minor activity. Therefore, for the decomposition of image sequences to be automated, much more research is needed even within the different major video classes.

4 Summary

The velocital information feature example demonstrates that video analysis tools are impacted by the video structure. The same video analysis tool is used on a typical cinematic sequence with shot boundaries and an air-to-ground airborne infrared sequence which contains erratic noise. The result shows that, in one case, the shot boundaries are detected and in the other, the erratic noise is detected. Image sequence quality impacts the ability to detect shot boundaries and can cause failure in automated analysis techniques. *A priori* knowledge of the video structure allows proper identification of shot boundary or erratic noise. Using the *a priori* knowledge that the infrared sequence is noisy suggests that the velocital information feature should be used to filter out the noisy image frames prior to sending the sequence to the general purpose shot boundary detector which may also use a velocital information as a video analysis tool.

Variations within a major video structure class can cause variations in the video analysis tool. Therefore, for the decomposition of image sequences to be automated, much more research is needed within the different major video classes.

This paper gave a key consideration for improving decomposition for digital image sequences for data storage and retrieval. The improvement is achieved by focusing on the specific digital image sequence structure. A cinemat-

ic video structure is compared to a surveillance video structure, and an airborne infrared video structure. The differences between the different digital video structures dictate different image sequence analysis approaches. The challenge for the storage and retrieval of digital video is to closely consider the video structure of the motion imagery.

References

[1] A. M. Tekalp. *Digital Video Processing.* Prentice Hall, Inc., Upper Saddle River, New Jersey, 1995

[2] C. A. Poynton. *A Technical Introduction to Digital Video.* John Wiley and Sons, Inc., New York, 1996.

[3] Gregory J. Power, Mohammad A. Karim, and Farid Ahmed. A Velocital Information Feature for Charting Spatio-Temporal Changes in Digital Image Sequences. *Journal of Electronic Imaging*, 8(2), 167–175, April, 1999.

Figure 9: Images from the sequence used for Figure 8 show the detected changes. From top to bottom the change is evident from frames 17 to 18 (orangutan resting, then gone from scene), 29 to 30 (gone, then sitting in scene), 49 to 50 (sitting, then interacting with visitor), and 58 to 59 (interacting with visitor, then visitor moves away).